G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model

Abstract

Large language models (LLMs) have shown remarkable proficiency in human-level reasoning and generation capabilities, which encourages extensive research on their application in mathematical problem solving. However, current work has been largely focused on text-based mathematical problems, with limited investigation in problems involving geometric information. Addressing this gap, we aim to enable LLMs to solve geometric problems by understanding image input. We first analyze the limitations of current Multimodal Large Language Models (MLLMs) in this area: they struggle to accurately comprehending basic geometric elements and their relationships. To overcome these challenges, we take advantage of the unique characteristics of geometric problems (such as unique geometric logical form, and geometric scalability) and the capacity of the textual LLMs to build an enriched multimodal geometry dataset based on existing data. The augmented dataset, Geo170K, contains more than 170K geometric image-caption and question-answer pairs. Utilizing our constructed Geo170K dataset, we develop G-LLaVA, which demonstrates exceptional performance in solving geometric problems, significantly outperforming GPT-4-V on the MathVista benchmark with only 7B parameters.